УДК 004.7:004.93:004.942

В.Г. Родченко, Е.В. Олизарович, А.И. Жукевич

ПРИМЕНЕНИЕ МЕТОДОВ РАСПОЗНАВАНИЯ ДЛЯ ПОСТРОЕНИЯ СИСТЕМ ДИАГНОСТИКИ КОМПЬЮТЕРНЫХ СЕТЕЙ

Предлагается универсальный метод диагностики компьютерных сетей, базирующийся на использовании математического аппарата теории распознавания образов. Представляется способ формализации состояния сети на основе наблюдаемых характеристик трафика. Рассматриваются механизмы реализации всех основных этапов диагностики, связанных с формированием словарей признаков, выполнением процедур обучения и классификации.

Введение

Компьютерные сети (КС) как объекты исследования имеют выраженные свойства сложных систем, в которых каждый компонент оказывает влияние на функционирование других. Современное состояние и тенденции развития телекоммуникационных сетей характеризуются распространением технологий мультисервисного доступа, беспроводных коммуникаций, виртуальных сетей, внедрением новых протоколов передачи данных и технологий распределенной обработки информации, повышением уровня компетентности пользователей. Эти процессы вызывают постоянное обновление сетевых устройств, программного обеспечения и характеристик сетевого трафика, что обусловливает появление качественно новых диагностических задач и приводит к необходимости постоянной модернизации методов и средств диагностики.

Понятия технической диагностики КС, определенные в нормативных документах, а также существующие технические средства и методы диагностики ориентированы на измерение и оценку значений среднестатистических показателей или точных значений (сигнатур), номенклатура которых постоянно возрастает и становится избыточной для конкретных задач управления. При этом анализ показателей возлагается на персонал, и результат существенно зависит от опыта и квалификации имеющихся специалистов. В целом средства диагностики КС ориентированы на определенный тип инцидентов, оборудования и программного обеспечения, что недостаточно в современных условиях.

Методы моделирования телетрафика, используемые при проектировании и теоретическом исследовании КС, также требуют дополнения, поскольку они, как правило, не могут непосредственно применяться для практической диагностики ввиду сложности их построения и невозможности учета всех необходимых данных. В связи с этим перспективным направлением является изучение динамически изменяющихся закономерностей функционирования телекоммуникационных сетей на основе апостериорного анализа свойств трафика гетерогенных КС с целью поиска новых методов моделирования и средств управления.

Статья посвящена описанию универсального метода диагностики КС, который основан на использовании математического аппарата распознавания образов. Реализованные механизмы диагностики инвариантны относительно технической структуры сети и ориентированы на автоматизацию операций мониторинга параметров информационных потоков и классификации состояния.

1. Постановка задачи

Комплексное решение задачи диагностики предполагает выполнение трех основных этапов, связанных, во-первых, с формализацией диагностируемых состояний, во-вторых, с реализацией процедуры обучения при построении образов эталонов диагностируемых состояний и, в-третьих, с постановкой заключительного диагноза, т. е. с классификацией состояния.

На первом этапе должны быть определены множество диагностируемых состояний (алфавит классов), а также набор наблюдаемых характеристик трафика и способы их формального

представления в виде априорного словаря признаков. В результате выполнения этого этапа должна быть сформирована классифицированная обучающая выборка (КОВ).

Второй этап связан с построением формальных образов эталонов состояний, компактных и разделенных в многомерном признаковом пространстве принятия решений. На основе анализа содержимого обучающей выборки выполняется процедура сепарирования признаков по степени их информативности с точки зрения разделения образов эталонов состояний, формируются уточненный словарь признаков и соответствующее пространство принятия решений, в котором размещаются образы эталонов состояний.

Заключительный этап предполагает выполнение процедуры классификации путем сопоставления образа наблюдаемого состояния КС и сформированного на предыдущем этапе множества образов эталонов состояний.

Распространенные средства диагностики КС обычно ориентированы только на решение третьего или, реже, второго этапа в виде программно-аппаратных комплексов, которые не могут быть модернизированы применительно к условиям конкретной КС. Такое положение позволяет эффективно решать распространенные известные диагностические задачи и исключать ошибки персонала, однако имеет и ряд недостатков:

1. Эталоны состояний, как правило, представляют собой одномерные величины, отражающие состояние одного фактора без анализа его взаимосвязи с другими, что не позволяет в полной мере осуществлять анализ состояния и прогноз развития сложных сетевых процессов.

2. Применение теоретически рассчитанных пороговых значений типовых параметров далеко не всегда позволяет оптимизировать структуру сложных сетей для решения специализированных задач диагностики в условиях конкретной сети.

3. Отсутствует возможность оперативного решения нестандартных эксплуатационных задач, анализа новых угроз, обнаружения аномальных состояний.

Таким образом, сегодня в процессе эксплуатации КС требуется не только оценивать соответствие известных параметров типовым значениям, но и проводить *анализ работы КС с целью обнаружения новых факторов или динамики изменения эталонных значений*. Существующие диагностические средства не предусматривают возможности решения таких задач.

Авторами разработан универсальный метод исследования и диагностики КС на основе анализа наблюдаемых характеристик трафика с использованием аппарата теории распознавания образов. Общая схема решения задачи диагностики состояний компьютерной сети формально может быть представлена в виде последовательности преобразований

$$S \longrightarrow C \longrightarrow A \longrightarrow T \longrightarrow A^{*} \longrightarrow E \longrightarrow R, \qquad (1)$$

где S – множество классов диагностируемых состояний, C – словарь наблюдаемых (измеримых) характеристик сети, A – априорный словарь признаков, T – классифицированная обучающая выборка, A^{*} – уточненный словарь признаков для построения пространства решений, E – множество эталонов диагностируемых классов состояний, R – множество решений.

Для реализации данной схемы в рамках настоящей работы исследованию подвергалась задача разработки и анализа необходимых для выполнения преобразований (1) алгоритмов:

– формализации состояния сети и построения априорного словаря признаков;

– формирования классифицированной обучающей выборки;

– сепарирования признаков из априорного словаря по степени их информативности для построения пространства решений;

– построения и представления образов эталонов диагностируемых состояний в пространстве решений;

– постановки заключительного диагноза.

Основой построения моделей и методов является применение аппарата теории распознавания образов и интеллектуального анализа данных [1–5].

2. Формализация состояний компьютерной сети

Процессы передачи данных, изначально инициированные прикладными программами, реализуются в виде потока пакетов в среде передачи. В настоящей работе принимается, что трафик КС как совокупность всех передаваемых пакетов является отражением технических и потребительских характеристик изучаемого сегмента сети. При исследовании различных аспектов работы КС под трафиком может пониматься: упорядоченный во времени поток кадров, поступающих на аппаратный порт; поток пакетов, проходящих через логический интерфейс; совокупность сообщений, полученных и отправленных прикладной программой.

В общем случае, поскольку передача данных в КС представляет собой многоуровневый процесс, каждое передаваемое по сети сообщение может быть формально описано в виде совокупности полей заголовков сетевых протоколов. Заголовки представляют собой форматированные области данных, включающие сведения об адресах источника и получателя сообщения, формате данных, длине сообщения, состоянии сеанса и т. д. Состав и структура информации, содержащейся в заголовках каждого уровня, известны и регламентируются спецификациями применяемых протоколов. Каждый передаваемый в КС кадр содержит информацию о всех сеансах, задействованных на разных уровнях в процессе взаимодействия клиента с сервером. Поэтому любой транспортный пакет, переданный через сегмент КС, можно описать конечным множеством стандартизированных параметров $\{x_1^{(1)}, ..., x_{j(i)}^{(i)}\}$, где $x_{j(i)}^{(i)}$ – значение *j*-го поля заголовка, предусмотренного на *i*-м уровне эталонной модели OSI/ISO. В предельном случае параметры $x_{j(i)}^{(i)}$ составляют универсальное множество – генеральную совокупность характеристик пакетов, включающую все возможные поля заголовков. Тогда каждый пакет данных в КС независимо от способа организации измерений может быть описан в виде вектора наблюдаемых характеристик:

$$\mathbf{x} = (x_1, ..., x_p), \tag{2}$$

где x_i – наблюдаемое (измеренное) значение характеристики $c_i \in C$, $i = \overline{1, p}$; p – количество характеристик кадра, измеримых программно-аппаратными средствами.

Для решения задачи диагностики целесообразно из всех возможных полей заголовка пакета выбрать только информативные, т. е. влияющие на качество классификации. Оценка информативности выполняется в два этапа:

1) *формализация* – эвристический выбор наиболее перспективных наблюдаемых (измеримых) параметров и построение на их основе априорного словаря для формального описания трафика;

2) *анализ* – исследование элементов априорного словаря на основе методов прикладной статистики.

Процессы передачи пакетов в реальной сети имеют, как правило, асинхронный и недетерминированный характер. Время возникновения пакета в канале передачи определяется как законами функционирования прикладных систем, обычно неизвестными наблюдателю, так и работой механизма доступа к среде передачи, результат работы которого в нагруженной сети также не может быть точно рассчитан. Таким образом, каждый наблюдаемый в сети пакет, формализованный в виде вектора $\mathbf{x}$, содержит информацию только об одном сеансе передачи и недостаточно информативен для диагностики состояния КС в целом.

Для практического применения данных, содержащихся в заголовках пакетов, требуется построить агрегированное описание трафика на основе фильтрации и селективного усиления значений наблюдаемых *характеристик пакетов* с целью получения формальных *признаков состояния A*. Для этого предлагается провести предварительную обработку данных о пакетах, зарегистрированных за определенный период измерений:

$$\mathbf{y} = (y_1, ..., y_n) = f(\mathbf{x}(t), \Delta t_x), \tag{3}$$

где y_j – значение признака состояния сети $a_j \in A$, $j = \overline{1, n}$; n – количество используемых для диагностики признаков; Δt_x – интервал агрегации; $\mathbf{x}(t)$ – значение вектора наблюдаемых характеристик, $t = [t_0, \Delta t_x)$, t_0 – начало интервала агрегации.

Величина Δt_x зависит от особенностей задачи и должна соответствовать интервалу наблюдения, в течение которого сеть проявляет все значимые для диагностики свойства. Для КС, находящейся в стационарном состоянии, результат диагностики для любого интервала Δt_x должен быть одинаков независимо от начала отсчета.

Длительность интервала Δt_x определяет оперативность выполнения диагностики, поэтому должна быть минимально достаточной для проявления трафиком своих свойств. Определение величины Δt_x может выполняться следующим образом: выбираться из типовых справочников, формироваться на основе требований диагностической задачи, выбираться на основе экспертных оценок, рассчитываться на основе предварительного анализа.

Словарь признаков $A=\{A_1, A_2,\ldots, A_n\}$ и соответствующие значения координат вектора **y** могут формироваться на основе следующих типовых преобразований [6]:

1. *Агрегация и фильтрация*:

– суммирование значений координаты x_i векторов $\mathbf{x}$, измеренных за период Δt_x:

$$y_j(\Delta t_x) = \sum_{q=1}^{m} x_{iq}, \tag{4}$$

где $j = \overline{1, n}$ – идентификатор признака состояния a_j; m – количество кадров, измеренных за период Δt_x; i – идентификатор используемой характеристики кадра c_i;

– подсчет количества событий вида $x_i = B$, зафиксированных за период Δt_x:

$$y_j(\Delta t_x) = \sum_{q=1}^{m} r, \text{ где } \begin{cases} r = 1 \text{ при } x_i = B; \\ r = 0 \text{ при } x_i \neq B, \end{cases} \tag{5}$$

где $j = \overline{1, n}$ – идентификатор признака состояния a_j; m – количество кадров, измеренных за период Δt_x; i – идентификатор используемой характеристики кадра c_i; B – заданная сигнатура;

– подсчет количества зафиксированных за период Δt_x случаев одновременного наступления событий $x_i = B_1$ и $x_k = B_2$, $i \neq k$:

$$y_j(\Delta t_x) = \sum_{q=1}^{m} r, \text{ где } \begin{cases} r = 1 \text{ при } x_i = B_1 \wedge x_k = B_2; \\ r = 0 \text{ при } x_i \neq B_1 \vee x_k \neq B_2, \end{cases} \tag{6}$$

где $j = \overline{1, n}$ – идентификатор признака состояния a_j; m – количество кадров, измеренных за период Δt_x; i, k – идентификаторы используемых характеристик кадра c_i, c_k; B_1, B_2 – заданные сигнатуры.

2. *Расчет статистических характеристик выборки k результатов измерений за период Δt_x:*

– выборочное среднее значение:

$$y_j(\Delta t_x) = \overline{x_i} = \frac{1}{k}\sum_{i=1}^{k} x_i; \tag{7}$$

– выборочная дисперсия:

$$y_j(\Delta t_x) = \mathrm{s}^2 = \frac{1}{k}\sum_{i=1}^{k} x^2{}_i - \left(\frac{1}{k}\sum_{i=1}^{k} x_i\right)^2. \tag{8}$$

Приведение к единому масштабу полученных значений параметров y_j обеспечивается нормировкой по диапазону разброса значений:

$$y^*_j = \frac{y_j - \min\limits_{j=\overline{1,m}}(y_j)}{\max\limits_{j=\overline{1,m}}(y_j) - \min\limits_{j=\overline{1,m}}(y_j)}, \quad (9)$$

где $j = \overline{1, n}$ – идентификатор признака состояния; m – объем выборки результатов измерений.

Вектор значений признаков **y**, полученный на основе вектора первичных характеристик **x** с использованием преобразований (4) – (9), представляет собой *формализованное описание сетевого трафика*, которое может быть использовано для анализа и диагностики компьютерных сетей. Применение такой формализации соответствует поставленным задачам и позволяет построить универсальный метод диагностики со следующими свойствами:

– независимость от средств измерений и возможность синтеза признаков на основе разнородных наблюдаемых данных;

– возможность получения новых знаний о свойствах КС на основе признаков, которые недоступны для методов прямых измерений;

– возможность предварительного преобразования первичных данных любого типа в порядковые величины, пригодные для сравнения и статистической обработки.

Предложенный в рамках настоящей статьи алгоритм формализации описания сетевого трафика (3) рассматривается как основа для построения *модели состояния сети* с использованием образов эталонных и диагностируемых состояний КС.

3. Построение пространства принятия решений и образов эталонов

Исследование и диагностика КС на основе использования аппарата теории распознавания и представленного выше алгоритма формализации предполагает решение следующих задач:

1. Формирование набора из k диагностируемых состояний и априорного словаря из n признаков состояния КС $A = \{A_1, A_2, \ldots, A_n\}$. Первичное формирование априорного словаря A может осуществляться либо на основе типовых библиотек, созданных на основе опыта решения подобных задач, либо на основе предположений экспертов.

2. Анализ признаков из априорного словаря и построение пространства принятия решений A^* для исключения неинформативных и неразделяющих признаков. Оценка информативности признаков априорного словаря выполняется с использованием классифицированной обучающей выборки T.

3. Построение множества многомерных эталонов состояния КС $E = \{E_1, E_2, \ldots, E_k\}$, где k – количество диагностируемых состояний.

4. Классификация исследуемого состояния в пространстве решений R.

Результативность предлагаемого метода диагностики КС возрастает с увеличением размерности априорного словаря признаков, поэтому на этапе формирования этого словаря необходимо стремиться найти наибольшее количество характеристик сети, логически связанных с диагностируемыми состояниями. Однако для повышения эффективности и снижения объема вычислений следует решить обратную задачу минимизации размерности признакового пространства. Последнее можно обеспечить путем определения признаков, наиболее информативных с точки зрения разделения эталонов состояний в пространстве решений. Для этого предлагается провести анализ информативности признаков априорного словаря на основе КОВ.

КОВ T строится на основе данных о значениях вектора **y**, наблюдаемых при нахождении компьютерной сети в каждом из i диагностируемых состояний. При этом каждое i-е состояние формально описывается в виде матрицы $T^{(i)}$, содержащей множество значений вектора $\mathbf{y}^{(i)}$:

$$T^{(i)} = \begin{pmatrix} y^{(i)}_{11} & y^{(i)}_{12} & \cdots & y^{(i)}_{1m_i} \\ y^{(i)}_{21} & y^{(i)}_{22} & \cdots & y^{(i)}_{2m_i} \\ \cdots & \cdots & \cdots & \cdots \\ y^{(i)}_{n1} & y^{(i)}_{n2} & \cdots & y^{(i)}_{nm_i} \end{pmatrix}, \qquad (10)$$

где i – идентификатор классифицируемого состояния; n – количество признаков в априорном словаре A; m_i – количество выполненных измерений сети, находящейся в i-м состоянии; $y^{(i)}_{jh}$ – значение j-го признака состояния $a_j \in A$, $j=\overline{1, n}$, полученное при h-м измерении, $h=\overline{1, m_i}$.

Анализ информативности признаков. В примененной схеме (1) для каждого из k диагностируемых состояний $s_i \in S$, где $i=\overline{1, k}$, КОВ определяет множество m_i объектов, которые на основе признаков из априорного словаря можно представить в виде вектор-столбцов вида $\mathbf{y}^{(i)T} = (y^{(i)}_1, y^{(i)}_2, \ldots, y^{(i)}_n)$ в n-мерном признаковом пространстве [7]. Объединение таких векторов для всех k состояний образует исходную КОВ в виде матрицы, содержащей n строк и m столбцов:

$$T = \begin{pmatrix} y^1_{11} & \cdots & y^1_{1m_1} & y^i_{11} & \cdots & y^i_{1m_i} & y^k_{11} & \cdots & y^k_{1m_k} \\ y^1_{21} & \cdots & y^1_{2m_1} & y^i_{21} & \cdots & y^i_{2m_i} & y^k_{21} & \cdots & y^k_{2m_k} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ y^1_{n1} & \cdots & y^1_{nm_1} & y^i_{n1} & \cdots & y^i_{nm_i} & y^k_{n1} & \cdots & y^k_{nm_k} \end{pmatrix}, \qquad (11)$$

где $m = m_1 + \ldots + m_k$; m_i – количество объектов i-го класса. При этом для каждого $s_i \in S$ можно выделить соответствующую матрицу размерности $n \times m_i$.

Для анализа информативности признаков предлагается выполнить их сепарирование на основе использования статистических критериев однородности. С этой целью для каждой пары состояний (1:2; 1:3; …; 1:k; 2:3; …; 2:k; …; (k−1):k) проверяется однородность множеств значений признаков из априорного словаря A – строк матрицы T. В качестве критериев однородности могут использоваться критерии Вилкоксона, Манна – Уитни и др. На основе анализа строк матрицы (11) в n-мерном множестве априорных признаков A можно выделить три подмножества $A^{(1)}$, $A^{(2)}$, $A^{(3)}$, для которых выполняется $A = A^{(1)} \cup A^{(2)} \cup A^{(3)}$, $n = n1+n2+n3$. Соответственно все признаки состояния a_j (где $j = \overline{1, n}$) могут быть разделены на три вида по следующему правилу:

– если для всех пар классов подтвердились гипотезы об однородности выборок значений j-го признака, признак a_j относится к *первому виду*;

– если для всех пар классов оказалось, что выборки значений признака a_j для двух сравниваемых классов подтвердили гипотезу об их статистической неоднородности, этот признак относится ко *второму виду*;

– если для признака a_j не выполнилось ни одно из двух предыдущих условий, он относится к признакам *третьего вида*.

Далее для построения модели состояния КС используются признаки второго вида. Если в результате сепарирования подмножество $A^{(2)}$ оказалось непустым (т. е. $n2 \neq 0$), то пространство принятия решений пригодно для достоверной классификации состояний [8]. Формируется уточненное пространство принятия решений $A^* = A^{(2)}$, а из матриц $T^{(i)}$ удаляются неиспользуемые строки. Если же подмножество $A^{(2)}$ оказалось пустым (т. е. $n2 = 0$), априорный словарь должен быть сформирован заново или требуется уточнение постановки задачи.

Результатами этапа анализа информативности признаков являются:

1. Обоснование факта возможности выполнения диагностики, т. е. существования эталонов диагностируемых состояний, которые формально представляются в виде компактных и разделенных образов в пространстве принятия решений.

2. Снижение размерности исходного множества признаков, построение пространства принятия решений A^*.

3. Формирование множества эталонов состояний $E = \{E_1, E_2, ..., E_k\}$, которые представляют собой классифицированные группы вектор-столбцов КОВ, полученные в результате сокращения априорного словаря признаков.

4. Классификация состояния

Для комплексного анализа состояния сложных информационно-технических сетей наиболее применимы методы кластерного анализа, которые базируются на использовании метрик расстояний между матрицами-эталонами E_i и матрицей диагностируемого состояния, рассматриваемой как дополнительный эталон E_d. Выбор метрик зависит от особенностей диагностической задачи. Могут быть использованы расстояние Хэмминга, евклидово расстояние, обобщенное расстояние Махалонобиса и др. Для оценки расстояния между кластерами в зависимости от задачи могут быть использованы следующие метрики: расстояние ближайшего соседа, расстояние дальнего соседа, расстояние между центрами тяжести, обобщенное расстояние по Колмогорову и др.

Каждый эталон вида (10) соответствует бесконечному множеству компактно размещенных объектов, обладающих общими свойствами и образующих *кластер*. Если такие эталонные кластеры не имеют пересекающихся областей, компактны и значительно «отдалены» друг от друга, то возможно упрощенное описание эталонного кластера на основе применения различных статистических оценок (математического ожидания, среднеквадратичного отклонения и т. д.).

В реальности исследователь обычно имеет дело с ограниченной выборкой измерений, качество которой существенно зависит от способа и условий измерений. Поэтому при диагностике состояний КС в многомерном пространстве признаков требуется использование эталонных кластеров сложной формы, точный расчет которых на практике невозможен в связи с объективным недостатком данных. В этих условиях для решения задач диагностики сети предлагается использовать метод переменных гиперсфер [9].

Для применения метода переменных гиперсфер исходными данными являются полученные в процессе анализа контрольной КС математические описания эталонов состояний сети E_i. Совокупность всех эталонов E_i образует матрицу вида

$$E = \begin{pmatrix} y^1_{11} & \cdots & y^1_{1m_1} & y^i_{11} & \cdots & y^i_{1m_i} & y^k_{11} & \cdots & y^k_{1m_k} \\ y^1_{21} & \cdots & y^1_{2m_1} & y^i_{21} & \cdots & y^i_{2m_i} & y^k_{21} & \cdots & y^k_{2m_k} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ y^1_{(n2)1} & \cdots & y^1_{(n2)m_1} & y^i_{(n2)1} & \cdots & y^i_{(n2)m_i} & y^k_{(n2)1} & \cdots & y^k_{(n2)m_k} \end{pmatrix}, \quad (12)$$

где i – идентификатор классифицируемого состояния; $n2$ – количество признаков в рабочем пространстве принятия решений A^*; m_i – количество выполненных измерений сети, находящейся в i-м состоянии; $y^{(i)}_{jh}$ – значение j-го признака состояния $a_j \in A$, $j=\overline{1, n}$, полученное при h-м измерении, $h = \overline{1, m_i}$.

Если строки матрицы заполнялись последовательно в процессе измерения, то каждый эталон E_i может рассматриваться как неупорядоченное в пространстве признаков множество $n2$-мерных векторов $\mathbf{y}^{(i)}_j$, $j = \overline{1, m_i}$, построенных в соответствии с моделью формализации (3).

Тогда каждому *i*-му эталону соответствует некоторая область пространства значений вектора наблюдения – кластер, все точки которого могут быть заданы по следующему алгоритму:

1. Построение гиперсфер для пары объектов. Выбирается произвольный объект $\mathbf{y}_1^{(i)}$ и строится множество расстояний $l^{(1)}_j$ до всех остальных объектов $\mathbf{y}_j^{(i)}$, $j \neq 1$. Находится ближайший объект $\mathbf{y}_2^{(i)}$, для которого выполняется $l^{(1)} = \min\limits_{j=2,\,m_i}(l_j^{(1)})$. Строятся сферы радиуса $r^{(1)}=\dfrac{l^{(1)}}{2}$ с центрами в вершинах векторов $y_1^{(i)}$, $y_2^{(i)}$. Обозначается точка касания двух сфер $\mathbf{g}^{(1)}$ с координатами $(g_1^{(1)}, ..., g_{(n2)}^{(1)})$ и строится сфера радиуса $r^{(1)}$ с центром в $\mathbf{g}^{(1)}$. Предположим, что полученные для каждой пары соседних объектов три пересекающиеся гиперсферы описывают кластер, все элементы которого считаются относящимися к *i*-му классу состояний сети. В общем случае для описания такого элементарного кластера G_i можно использовать следующее представление:

$$G_i = \begin{pmatrix} \mathbf{y}_j^{(i)} & r & b \\ \mathbf{g}^{(j)} & r^{(j)} & b \\ \mathbf{y}_{(j+1)}^{(i)} & r & b \end{pmatrix}, \tag{13}$$

где $j = \overline{1, m_i}$; m_i – количество векторов $\mathbf{y}^{(i)}$ объектов в исходном эталоне E_i; $b = 1$, если центром сферы является реальный вектор $\mathbf{y}$; $b = 0$, если центром сферы является точка $\mathbf{g}$; $r = r^{(j)}$ для объектов с номерами $j = 1$ и $j = m_i$ (крайние точки эталона); величина $r \in [\min(r^{(j)}, r^{(j+1)}), \max(r^{(j)}, r^{(j+1)})]$ для остальных точек эталона зависит от постановки задачи диагностики.

2. Упорядочивание множества эталонных объектов. Аналогично (14) строятся элементарные кластеры для остальных пар ближайших объектов эталона данного класса E_i и формируется общее описание эталонного кластера *i*-го класса в виде упорядоченного множества гиперсфер:

$$E_i \Rightarrow G_i = \begin{pmatrix} 1 & \mathbf{y}_1^{(i)} & r^{(1)} & b \\ 2 & \mathbf{g}^{(1)} & r^{(1)} & b \\ 3 & \mathbf{y}_2^{(i)} & r & b \\ \ldots & \ldots & \ldots & \ldots \\ 2m_i - 2 & \mathbf{g}^{(m_i-1)} & r^{(m_i-1)} & b \\ 2m_i - 1 & \mathbf{y}_{m_i}^{(i)} & r^{(m_i-1)} & b \end{pmatrix}. \tag{14}$$

3. Классификация состояния на основе эталонных кластеров. Исследуемое состояние сети может быть задано в виде одного или нескольких значений вектора $\mathbf{y}^{(d)}$, которые образуют эталон классифицируемого состояния E_d. Тогда для решения задачи классификации состояния необходимо рассчитать «меру близости» между каждым *i*-м эталонным кластером $G_i : E_i \Rightarrow G_i$ и контрольным диагностируемым кластером $G_d : E_d \Rightarrow G_d$.

5. Применение метода

Рассмотренный метод диагностики разработан для автоматизации прикладных исследований КС и основан на универсальном наборе алгоритмов моделирования трафика, верификации значимости признаков и классификации состояний. Метод ориентирован на использование в случаях, когда прямые измерения характеристик сети невозможны или недостаточно информативны и строгие математические правила классификации отсутствуют. Способ извлечения

информации об объекте на основе анализа косвенных признаков широко применяется в геологии, радиолокации, медицине и других отраслях.

Распространенные методы анализа состояния сетей, в том числе компьютерных, основаны на общих для диагностики технических систем принципах, таких как измерение стандартных характеристик и выявление их текущих отклонений от нормального состояния на основе одномерных пороговых значений и сигнатур (HP OpenView, NetIQ, SolarWinds, Fluke, Snort, Bro и др.). Многокритериальный анализ КС с использованием таких систем основан на алгоритмах последовательного анализа, предполагающих поочередный расчет параметров и их иерархическую обработку [10, 11]. Данный подход дает хороший результат для поиска неисправностей технических подсистем, но менее эффективен при анализе информационных процессов, поскольку требует априорного знания набора параметров и их допустимых значений, а результат может зависеть от очередности выполняемых операций.

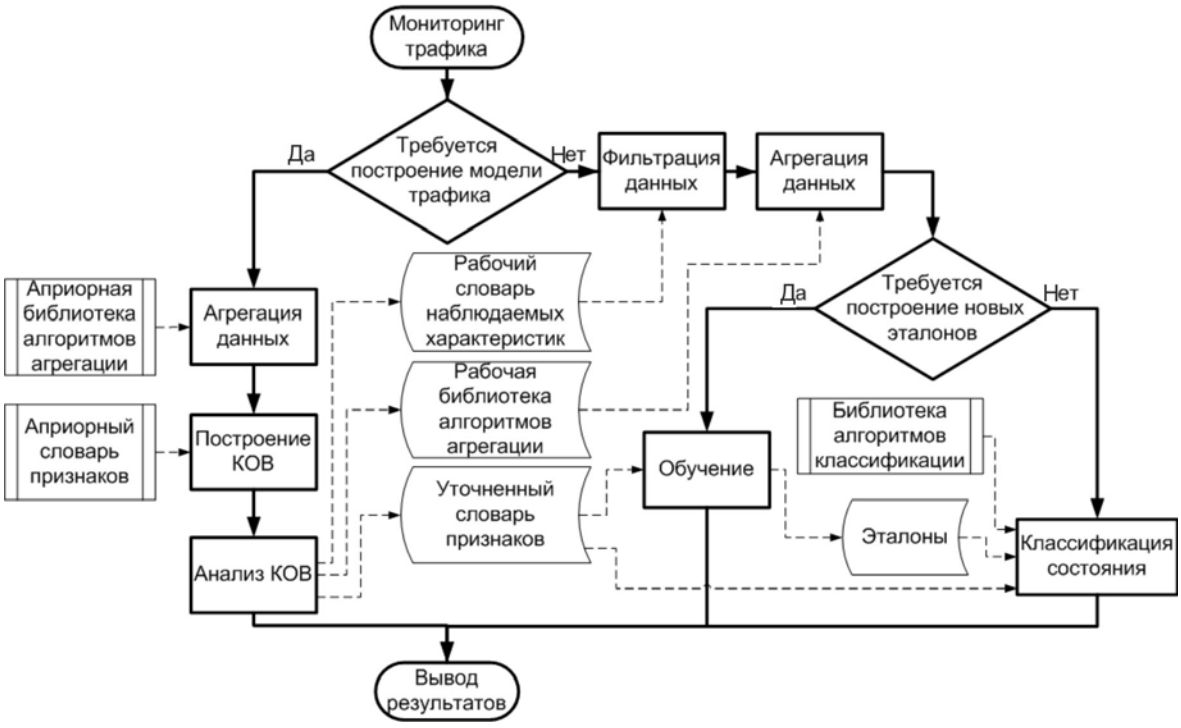
Предлагаемый авторами метод позволяет использовать результаты работы указанных и любых других диагностических средств в качестве первичных данных о трафике и на их основе выполнять не только оценку текущего состояния, но и анализ долгосрочных динамических процессов в сети. При этом набор параметров КС может формироваться и верифицироваться самой диагностической системой, а за счет применения алгоритмов кластеризации появляется возможность анализа многомерных данных, что упрощает автоматизацию процесса и позволяет получить «объемный» срез событий в КС. Таким образом, в настоящей работе под диагностикой понимается не констатация технического состояния отдельных подсистем и узлов, а комплексная оценка и прогнозирование развития сети.

Диагностическая система, основанная на предложенном методе, позволяет решать недоступные ранее классы задач, но является более сложной для эксплуатации и требует не только технической, но и математической подготовки персонала при реализации трех основных режимов работы (рисунок):

исследование – режим, позволяющий строить и верифицировать новые модели трафика, рабочие словари, библиотеки алгоритмов;

обучение – режим, позволяющий на основе существующего рабочего словаря строить базовые эталоны состояний для режима классификации сети;

классификация – режим, соответствующий режиму «диагностика» в стандартных системах.



Режимы функционирования системы диагностики

В основе работы такой диагностической системы лежат программные модули и множества данных (априорные и рабочие словари, эталоны, библиотеки алгоритмов), являющиеся в общем случае многомерными и сформированные с помощью последовательного выполнения всех режимов или готовых внешних библиотек. В статье рассмотрен один набор математических методов, но при построении диагностических систем различного назначения могут независимо изменяться первичные источники данных о трафике, алгоритмы агрегации, методы исследования однородности выборки, способы построении кластеров и оценки расстояний между кластерами.

В отличие от традиционных способов формализации рассмотренный метод диагностики использует не сами измеряемые характеристики, а их производные, что позволяет конвертировать любые типы данных в порядковый тип, пригодный для автоматизированной компьютерной обработки. В основе метода лежит предложенный способ формализации трафика (2) – (9), позволяющий учесть и использовать особенности сетевых структур, такие как множественность топологических и функциональных связей и высокую скорость протекания процессов. В качестве критериев классификации могут применяться как пороговые значения межкластерных расстояний, так и многомерные сигнатуры.

На основе знаний о многоуровневой структуре трафика КС и анализе его многомерных моделей изложенный в работе метод диагностики позволяет решать следующие прикладные задачи, требующие исследования индивидуальных свойств сети:

– поиск зависимостей в работе сегментов, устройств и сервисов сети на основе создания индивидуальных правил;

– обнаружение скрытых и аномальных явлений и процессов;

– обнаружение «медленных» изменений и построение прогнозов функционирования КС;

– оценка характера использования шифрованных соединений (p2p, VPN);

– поиск «невидимых» в сети устройств (шлюзов, неуправляемых коммутаторов, мостов);

– поиск сигнатур и сенсоров для других диагностических систем (Snort, Cisco ASA и др.).

К недостаткам метода можно отнести более низкую в общем случае достоверность результатов, обусловленную применением методов распознавания. В проведенных авторами опытах по обнаружению в сети узлов специального типа (серверов, маршрутизаторов, сканеров) получен уровень достоверности 70–90 % при применении эталонов с шестью показателями. При использовании одномерных показателей и стандартных диагностических задач система функционирует аналогично другим средствам.

Применение рассмотренного метода наиболее оправдано в следующих случаях:

– при избирательной диагностике событий, характерных и критичных для бизнес-процессов в данной сетевой системе;

– диагностике гетерогенных сетей с разнородными, в том числе неуправляемыми сетевыми устройствами;

– необходимости автоматизации принятия решений о состоянии сети.

Сложность практического применения метода состоит в его ориентации на индивидуальные свойства сети, т. е. максимальная эффективность достигается, если система диагностики реализована не в виде готового измерителя, а в виде набора настраиваемых математических инструментов, что актуально, например, для организаций, специализирующихся на аудите компьютерных систем.

Заключение

В работе предложен метод диагностики комплексных состояний компьютерной сети, основанный на апостериорном исследовании трафика КС и анализе его статистических характеристик с применением алгоритмов теории распознавания образов. Представлена реализация основных этапов построения системы диагностики компьютерной сети: формализация состояния сети, анализ информативности признаков и формирование пространства решений, построение эталонов, классификация состояний.

Разработанная модель диагностики применима для решения задач различного уровня, в том числе она может быть адаптирована к уже существующим моделям (контроль пороговых значений, поиск сигнатур и др.). В качестве источников первичных данных могут использо-

ваться любые программно-аппаратные системы журналирования сетевой активности: анализаторы сети, снифферы и сканеры, подсистемы аудита системных событий, log-файлы серверных программ, а также результаты работы других систем диагностики.

Предложенный метод предназначен для построения автоматизированных систем диагностики, позволяющих перенести значительную часть аналитической работы на программно-технические системы, так как все этапы решения диагностической задачи формализованы и решаются с использованием методов и алгоритмов теории распознавания образов и кластерного анализа.

Перспективными направлениями развития рассмотренного метода являются: построение динамических моделей для прогнозирования работы КС на основе обработки уже полученных результатов диагностики, создание библиотек словарей признаков и эталонов для типовых задач диагностики, исследование и оптимизация интервалов агрегации для разных задач, совершенствование методов точного построения эталонов состояний.

Список литературы

1. Олизарович, Е.В. Метод и технология построения систем диагностики компьютерных сетей на основе распознавания образов : автореф. дис. … канд. техн. наук : 05.13.13 / Е.В. Олизарович. – Гродно : ГрГУ, 2010. – 23 с.

2. Загоруйко, Н.Г. Прикладные методы анализа данных и знаний / Н.Г. Загоруйко. – Новосибирск : Изд-во Института математики СО РАН, 1999. – 268 с.

3. Айвазян, С.А. Прикладная статистика: основы моделирования и первичная обработка данных : справ. изд. / С.А. Айвазян. – М. : Финансы и статистика, 1983. – 471 с.

4. Горелик, А.Л. Методы распознавания : учеб. пособие для вузов / А.Л. Айвазян. – 3-е изд. – М. : Высш. шк., 1989. – 232 с.

5. Журавлев, Ю.И. Распознавание. Математические методы. Программная система. Практические применения / Ю.И. Журавлев. – М. : Фазис, 2005. – 159 с.

6. Олизарович, Е.В. О формализации описания компьютерной сети для диагностики на основе методов распознавания / Е.В. Олизарович, В.Г. Родченко // Информационные системы и технологии (IST'2009) : материалы V Междунар. конф.-форума, Минск, 16-17 ноября 2009 г. – Минск : А.Н. Вараксин, 2009. – С. 164–167.

7. Жукевич, А.И. Метод построения эталонов состояний компьютерной сети на основе применения алгоритмов теории распознавания образов / А.И. Жукевич, Е.В. Олизарович, В.Г. Родченко // Сетевые компьютерные технологии : сб. тр. III Междунар. науч. конф., 17–19 окт. 2007 г. – Минск : Изд. центр БГУ, 2007. – С.14–17.

8. Жукевич, А.И. Об одном методе построения компьютерной системы диагностики состояний технологических процессов / А.И. Жукевич, Е.В. Олизарович, В.Г. Родченко // VIII Междунар. конф. «Интеллектуальный анализ информации ИАИ-2008», Киев, 14-17 мая 2008 г. – Киев : Просвіта, 2008. – С. 398–406.

9. Жукевич, А.И. Использование метода Монте-Карло для реализации процедуры вычисления объема пересечения сфер в пространстве R^n / А.И. Жукевич, В.Г. Родченко // Актуальные проблемы математики и компьютерного моделирования : сб. науч. ст. – Гродно : ГрГУ, 2007. – С. 192–196.

10. Бредихин, С.В. Идентификация сканеров в IP-сетях статистическим методом / С.В. Бредихин, В.И. Костин, Н.Г. Щербакова // Проблемы информатики. – 2008. – № 1. – С. 22–36.

11. Хогдал, Дж. Скотт. Анализ и диагностика компьютерных сетей / Дж. Скотт Хогдал. – М. : Лори, 2001. – 360 с.

Поступила 27.06.2012

Гродненский государственный университет им. Я. Купалы,
Гродно, ул. Ожешко, 22
e-mail: rovar@mail.ru, e.olizarovich@grsu.by, san@grsu.by

V.G. Rodchenko, E.V. Olizarovich, A.I. Zhukevich

PATTERN RECOGNITION METHODS FOR BUILDING A NETWORK DIAGNOSTIC SYSTEM

A generic method for computer networks diagnosis is proposed, which is based on the theory of pattern recognition. We suggest to model the network status on the basis of the observed traffic's characteristics. Implementation mechanisms of all diagnostic stages are considered. They include formation of the dictionaries and performing procedures of training and classification.